Research Data, Big Data, and Chemistry

by Richard Hartshorn

The IUPAC centenary in 2019 is fast approaching, and this will naturally lead people to look back at the significant achievements of the organisation and its dedicated volunteers over the past one hundred years. Equally important, however, will be the need to look forward to the roles for IUPAC in its second century. This special issue of Chemistry International (CI) could well feature in that assessment, as technology in the digital age, and particularly the data that technology produces, will clearly be an essential tool for the future of chemistry as a discipline.

The IUPAC vision, as espoused in our new strategic plan, is to be an indispensable resource for chemistry through the development of tools for the application and communication of chemical knowledge. In this issue of *CI*, you will find examples of the ways that data analysis can assist chemists and lead to the evolution of new chemical knowledge, and also of the ways that the effective utilization of data can assist in the communication of that knowledge.

Throughout its nearly-completed first century, IUPAC has been recognised particularly for its contributions to nomenclature, terminology, and the symbols of chemistry; for its standardisation of chemical methods; and for its critical evaluation of data and the development of standards for data exchange. The colour books and the curation of the periodic table, along with the atomic weight data within it, are particularly well-known, widely-used, and appreciated by students and researchers alike—even if they may sometimes appear to be a necessary evil. The periodic table and atomic weight data will always be essential to the discipline, and some of the uses to which it has been put are both fascinating and educational. [1] By contrast, many people have commented on the reduced importance of conventional nomenclature (and by implication, the colour books), as the quality of structure drawings and the ease with which such drawings can be incorporated into documents, websites, and other media has improved. Indeed, the rise of the graphical representation of molecules in documents has created challenges for database manipulation and searchability, and it is within this context that the IUPAC International Chemical Identifier (InChI) was invented, implemented, and developed. [2] The InChI identifier is now globally embraced and is being used in a wide variety of applications. In fact, in this issue of *CI* you will find InChI mentioned numerous times under a variety of topics. This issue of *CI* will also address a wide-range of issues in data management and data usage across the entire discipline.

From a personal perspective, my involvement with IUPAC mirrors, at least in a small way, the evolution of IUPAC activity over recent times. I began as part of the team producing a new version of the "Red Book", *Nomenclature of Inorganic Chemistry–IUPAC Recommendations 2005* [3] and then took on leadership roles in the Division of Chemical Nomenclature and Structure Representation (Division VIII). In those roles I was involved in the development of standards for graphical representation, [4,5] which collectively were guides to drawing chemical structure diagrams that are as unambiguous and informative as possible. I also began to learn more about InChI, particularly about its use in database management/merging. It gradually became clear to me that there was potential for significant application of InChI's beyond databases, and I have become involved in InChI development, at least in a small way, through projects on the development of InChI QR codes [6] and InChI for mixtures. [7]

Now, in my role as IUPAC Secretary General, one of my major responsibilities is to help identify and encourage the development of new IUPAC activities and projects, particularly those that have strategic importance: those that will shape future IUPAC activities and enhance IUPAC's relevance in its second century. One of the key steps in doing this is to collaborate with other organisations and groups that have similar interests. I have been very pleased to see the development of collaborations between the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS) and the Chemistry Interest Group of the Research Data Alliance (RDA) and those individuals and organisations who are involved with it.

This special issue of *CI* describes many of the recent activities that I believe will have future significance, given the likely importance of "Big Data," the potential of data mining, and the benefits that will derive from being able to properly search, access, and mine all of the research data that scientists around the globe are busily accumulating.

(continued on page 4)

The Rise of Primary Research Data

by Leah McEwen and David Martinsen

As the scale of global commerce and opportunities for multidisciplinary collaboration increase, there is greater pressure on basic research to supply a quick return on investment (ROI). The emergence and development of digital information technologies in the new millennium have inspired a new look at how research outputs are managed and disseminated. The driving question in the minds of many research funders is this—will lowering the barriers for access increase the value of research for the greater society? This is a particularly interesting question to consider for measurement data, the greater amount of which are scattered across millions of separate, fixed publications (not to mention those never published and lingering in file drawers and on hard drives). Can the advent of cloud technologies, exchange standards, and provenance tracking facilitate improved access, evaluation, and use of data for both research and commerce? Can new value and discovery be realized through the greater aggregation of measured scientific data as "Big Data"?

The past five years has seen practical conversations among stakeholders increasingly focused on the publication of primary research data associated with journal articles. Data publication advocates have lobbied for the availability of data, funding agencies have issued mandates requiring funded researchers to publish their data, and repositories have been created to support researchers in fulfilling these requirements. The arguments put forth are many: it is important that science be as transparent as possible so that the community can properly assess the integrity of the research being published; it is valuable for interested scientists to have access to machine-readable data to more deeply examine and interact with the data described in a journal article; it is important that editors and reviewers have access to all of the available material to better understand the validity of the conclusions being presented, or consider whether the data themselves exhibit evidence of manipulation in a fraudulent manner.

This interest in the publication of research data, among other scholarly communication challenges, has spawned a number of new organizations (for example, FORCE11, [1] the Research Data Alliance), [2] which augment long-standing organizations (such as

The guest editors at the University of Tokyo.

CODATA [3] and ICSU [4]). In addition, repositories for depositing research datasets, such as Data Dryad, [5] figshare, [6] and Mendeley Data, [7] have appeared. In chemistry, these new services may, in some sense, augment traditional curated data collections, such as the former Beilstein and Gmelin Handbooks, the Cambridge Structural Database, [8] the Protein Data Bank, [9] the Powder Diffraction File, [10] the Spectral Database for Organic Compounds (SDBS), [11] Wiley and NIST's Mass Spectral Databases, [12,13] BioRad's Spectroscopy Databases, [14] and others.

As a result of the emerging expectations for researchers to publish data, scientific publishers and research libraries are beginning to offer support services to their communities in navigating this evolving landscape. Balancing both sides of the time-cost equation for data generators and consumers will be key to how well new practices are established.

Taking a look at how the movement to publish research data more accessibly intersects the practice of research data dissemination in chemistry is the impetus behind a **Special Symposium on Research Data, Big Data, and Chemistry** at the 46th IUPAC World Congress, and the basis for this special issue of Chemistry International. The perspectives represented here examine a range of issues from coordinating global initiatives to workflows for publication, review, and evaluation to education to applications in industry and

society. Also considered are some IUPAC digital initiatives for supporting chemistry data publication, including the International Chemical Identifier (InChI) [15] and the online Gold Book Compendium of Chemical Terminology. [16]

We hope you enjoy the reading, and look forward to meeting you at the Congress in São Paulo, Brazil, 9-14 July and the Special Symposium on 13 July 2017. [17]

References

1. www.force11.org
2. www.rd-alliance.org
3. www.codata.org
4. www.icsu.org
5. http://datadryad.org
6. https://figshare.com
7. https://data.mendeley.com
8. www.ccdc.cam.ac.uk
9. www.rcsb.org/pdb/home/home.do
10. www.icdd.com
11. http://sdbs.db.aist.go.jp
12. www.wiley.com/WileyCDA/WileyTitle/productCd-1119171016.html
13. http://chemdata.nist.gov
14. www.bio-rad.com/en-us/spectroscopy
15. http://www.inchi-trust.org/
16. https://goldbook.iupac.org/
17. www.iupac2017.org/special-symposia.php#tab2

Leah McEwen <lrm1@cornell.edu> chemistry librarian at Cornell University, USA. She is a member of the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS), and co-chair of the CPCDS Subcommittee on Cheminformatics Data Standards. ORCID.org/0000-0003-2968-1674

David Martinsen <dmartinsen@consultdpm.com>, formerly a Senior Scientist at ACS Publications, consults in scholarly publishing at David Martinsen Consulting in Rockville, MD, USA. He is co-chair of the CPCDS Subcommittee on Cheminformatics Data Standards, and is also a co-chair of the Research Data Alliance Chemistry Research Data Interest Group. ORCID.org/0000-0002-8667-5855

(continued from page 2)

References

1. www.isotopesmatter.com
2. www.inchi-trust.org
3. N. G. Connelly, T. Damhus, R. M. Hartshorn, A. T. Hutton, *Nomenclature of Inorganic Chemistry*, Royal Society of Chemistry, ISBN 0-85404-438-8, 2005.
4. J. Brecher, K. N. Degtyarenko, H, Gottlieb, R. M. Hartshorn, G. P. Moss, P. Murray-Rust, J. Nyitrai, W. Powell, A. Smith, S. Stein, K. Taylor, W. Town, A. Williams, A. Yerin, "Graphical Representation of Stereochemical Configuration", *Pure Appl. Chem.* **78**(10):1897-1970, 2006. https://doi.org/10.1351/pac200678101897
5. J. Brecher, K. N. Degtyarenko, H, Gottlieb, R. M. Hartshorn, K.-H. Hellwich, J. Kahovec, G. P. Moss, A. McNaught, J. Nyitrai, W. Powell, A. Smith, K. Taylor, W. Town, A. Williams, A. Yerin, "Graphical Representation Standards for Chemical Structure Diagrams", *Pure Appl. Chem.* **80**(2):277-410, 2008. https://doi.org/10.1351/pac200880020277
6. https://iupac.org/project/2015-019-2-800
7. https://iupac.org/project/2015-025-4-800

Richard Hartshorn <richard.hartshorn@canterbury.ac.nz> is a member of the chemistry faculty at the University of Canterbury, in Christchurch, New Zealand. So far, his involvement with IUPAC has been largely based in nomenclature, and dates from the late 1990s, when he was persuaded to join the group preparing a revision of the Red Book ("Nomenclature of Inorganic Chemistry, IUPAC Recommendations 2005", ISBN 0-85404-438-8). Since then he has been involved in numerous projects and has been a member of the Committee on Chemistry Education (since 2006). He was elected to positions of responsibility in the Division of Chemical Nomenclature and Structure Representation (Titular member 2003-07, Vice President 2008-09, President 2010-13) and the Bureau (2014-17), and took over as as IUPAC Secretary General in January 2016.